Applying Linked Data Technologies in the Social Sciences

Zapilko, Benjamin; Schaible, Johann; Wandhofer, Timo; Mutschke, Peter

Postprint / Postprint
Zeitschriftenartikel / journal article

Post-Print der Veröffentlichungsversion, erschienen in:
Künstliche Intelligenz, Online First, doi: 10.1007/s13218-015-0416-6

Applying Linked Data Technologies in the Social Sciences

Benjamin Zapilko[1] • Johann Schaible[1] • Timo Wandhofer[1] • Peter Mutschke[1]

benjamin.zapilko@gesis.org – [1]

GESIS - Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, 50667 Cologne, Germany

Abstract In recent years. Linked Open Data (LOD) has matured and gained acceptance across various communities and domains. Large potential of Linked Data technologies is seen for an application in scientific disciplines. In this article, we present use cases and applications for an application of Linked Data in the social sciences. They focus on (a) interlinking domain-specific information, and (b) linking social science data to external LOD sources (e.g. authority data) from other domains. However, several technical and research challenges arise, when applying Linked Data technologies to a scientific domain with its specific data, information needs and use cases. We discuss these challenges and show how they can be addressed.

1 Introduction

In recent years. Semantic Web standards and technologies have matured. In particular. Linked Open Data (LOD) [1] has gained popularity and acceptance across various communities and domains. It has encouraged numerous research organizations, archives, libraries, and governmental agencies to publish their data on the web. Science politics see a large potential for applying semantic technologies and Linked Data in scientific disciplines [6, 11], e.g. enhancing services and infrastructures providing research information. In order to investigate these potentials in the scientific domain of the social sciences, expert interviews with social scientists have been conducted in [13]. The results of the interviews show that there is currently no consumption of Linked Data in that particular domain, because there are no adequate applications for users. However, the interviewees can imagine using Linked Data technologies for linking, matching and enriching heterogeneous data and their collections. This may influence the search for complex information needs and the analysis of combined data sets.

In this article, we present use cases and challenges that arise when applying Linked Data technologies in the social sciences. The use cases concentrate on (a) linking data of different information types in a closed domain for allowing users to find interlinked domain-specific information, and on (b) linking the domain-specific data to external LOD sources (e.g. authority data) for providing the user with additional information. We show how these use cases are currently addressed in ongoing projects and applications. Finally, we provide a brief overview on challenges that arise when applying Linked Data technologies to a particular domain like the social sciences with its specific data, users, and their information needs.

2 Use Cases

In the following, we describe the two use cases for applying Linked Data in the social sciences and how they can be realized. Furthermore, we present two example applications that illustrate a visionary use of Linked Data in that domain.

Hereby, we distinguish between three actors: (1) the *data provider*, e.g. a research infrastructure organisation like GESIS, (2) the *application developer*, who integrates Linked Data in an application, e.g. a web portal, and (3) a *social science researcher*, who is a user of the application and has different information needs. However, although focusing on the social sciences, the presented use cases are representative for other scientific

[1] http://infolis.github.io/ (accessed 15/05/2015).

disciplines.

2.1 Interlinking Domain-Specific Information

When seeking research data in a web portal, researchers are often interested in information on whether and in which publications particular data sets have already been cited and analysed. Typically, this information is not explicitly available in the metadata of such data sets, since literature and research data are usually collected separately, in many cases also by different organizations.

Therefore, additional relationships are necessary that link information from different sources together in order to support searches across different information objects, such as research projects, literature and research data, or to enhance information aggregation, such as compiling all relevant information about a particular scientist into one integrated information object as shown in Fig. 1.

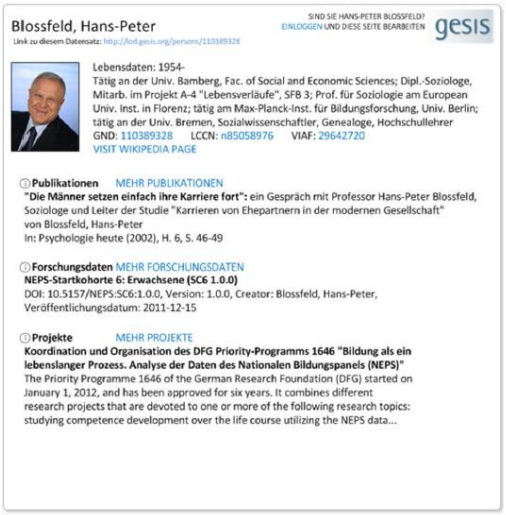

Fig. 1 Webpage with aggregated interlinked information

In order to present such relationships to users, links between records of different information types or between different data collections must be detected and made available by the data provider. Subsequently, application developers can include this information in web portals.

An example for that is the DFG-funded project InFoLiS at GESIS which aims at detecting links between literature and research data sets. This is performed by an iterative pattern induction method which recognizes references of data sets in full texts [2]. The approach achieves results with a very high precision, but requires only minimal supervision and shallow features (e.g. no layout information of the analysed pdf document is necessary).

In order to provide the links generated by this method to other services in a standardized way, they are stored in a RDF triple store. In a federated approach, the remaining data from records of different data collections can stay in the particular data collections and can be published as Linked Data on-the-fly. This approach allows a lightweight strategy for information integration instead of using an organization-wide, large relational database model which would not be scalable at a particular size and dynamics of change, in particular with respect to the circumstance that many data collections emerge and evolute independently from each other due to different project contexts. The described approach is also applicable in other scientific domains. For link detection like it is conducted in InFoLiS, it is then necessary to consider that citations between literature and research data may follow different patterns.

2.2 Linking Social Science Data to External LOD Sources

Another LOD use case is to link domain specific data to external data that is out of the scope of the domain under study. Examples for this are geographical information about places mentioned in data and literature, authority data on person names, or further explanations, governmental data and news headlines to a particular topic.

A concrete example for the use of external LOD sources is the inclusion of data from authority files like those of the German National Library^ or VIAF in order to precisely distinguish names of persons or organizations as shown in Fig. 2. Authority files contain unique identifiers for e.g. persons and organizations in order to reference these entities consistently, uniquely, and unambiguously.

Aligning person names by authority data may improve the precision of searches for persons immensely, in particular in cross-collection search scenarios. However, this assumption still needs to be verified.

Fig. 2 Distinguishing author names during the data collection process

One example application for this use case is the Semantic Data Library for the social sciences [4]. In this domain, it is a common research task to match heterogeneous research data sets (e.g. statistical and survey data) in order to analyse the combined data set. The developed prototype[2] served as an initial feasibility study and provides central services for the accessing, processing and integration of distributed LOD sources. Users were able to select two different data sets out of election statistics and well-being survey data and choose which particular parts should be presented in an integrated visualization. The physical storage location of the data sets remains distributed and is not collected or hosted by the data library. The difficulties in searching, modelling, and annotating distributed data are addressed not only on the metadata level, but also on the underlying numerical data. This provides researchers with on-the-fly usage of the data in visualizations or for statistical analysis.

Another example application aims at supporting policy makers with relevant LOD sources. In the EU project Sense4us[3], the policy-making process is supported by creating a knowledge map of internet-based information for a particular topic. Model building and simulation technologies enable the user to start with one particular keyword (e.g. renewable energy) to create this map that can include information from Twitter and LOD sources.

In the project, it turned out that many data sources relevant for policy making are not available as LOD so far. Thus, this data cannot be integrated in the knowledge map. Therefore, a prototype that converts Open Data sources into RDF and interlinks them to the LOD cloud[4] is currently under development.

3 Challenges

Along with the presented use cases, several challenges arise. In the following paragraphs, we discuss these challenges and describe how they can be addressed.

Data Modelling When modelling Linked Data, it is seen as best practice to reuse terms of existing vocabularies as far as possible instead of defining new ones [5]. However in [8], it was investigated that the reuse of existing vocabularies depends on the use case in which the data is being published as Linked Data. In most cases, it is desired to keep the published Linked Data as simple as possible in order to enable an easy processing by further applications. For this purpose, a vocabulary term recommendation tool can support the Linked Data engineer, who is often closely connected to a data publisher, in order to create a highly compatible and reusable Linked Data set. This tool can analyse vocabulary terms being used in other Linked Data sources and generates recommendations of suitable terms for the data that is being modelled by the engineer.

For modelling social science data as Linked Data, existing vocabularies and ontologies are not sufficient and expressive enough [13], e.g. for representing person-level data (like survey data). Hence, new vocabularies and extensions of existing vocabularies have become necessary. In the social, behavioural and economic sciences, DDI (Data Documentation Initiative)[7] is a widely accepted XML data model for representing metadata of person-level data. In order to represent this metadata in its full extent as Linked Data, the DDI-RDF Discovery vocabulary was developed [3]. For connecting research data in general with other information objects of the research process, it was included into the existing SWRC ontology [9], which allows for representation of research communities and activities [13]. Works presented in [12] allow for an extended SKOS representation of complex domain-specific thesauri holding specific term relationships that are originally not covered by SKOS like compound equivalences of terms. These previous efforts enable a complete publication of all data

[2] The prototype was the result of a cooperation between Karlsruhe Institute of Technology (KIT), the Institute for Web Science and Technologies (WeST), the statistical office and IT service provider of the federal state North Rhine-Westphalia IT.NRW and GESIS.

[3] http://www.sense4us.eu/ (accessed 15/05/2015). The project has received funding from the European Unions Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 611242.

[4] http://lod-cloud.net/ (accessed 15/05/2015).

[2] http://www.dnb.de/gnd (accessed 15/05/2015).

[3] http://viaf.org/viaf/data/ (accessed 15/05/2015).

that is relevant in the research context of the social sciences.

Data Linking In context of linking different data sets, two challenges arise: entity disambiguation and specification of links and their semantics. While there exist state of the art tools for link detection like Silk [10] or Limes [7], we have to investigate strategies for the disambiguation of these entities inside a single database and between different databases. Considering the linking of person names and organization names to authority files, we have to face the challenge of entity disambiguation. In the InFoLiS project, detected links between records of different information types are unspecified, i.e. it remains unclear whether the detected link between a research data set and a publication is a citation or whether the data set has been the analysed data source in the publication. However, a precise distinction between different semantics of links is necessary in order to use this link semantics for further applications or services. We plan to focus our further research on specifying and classifying links.

4 Conclusion

Applying Linked Data technologies in a particular scientific domain is not a straight forward process as our efforts have shown to date. Technical and research challenges often arise only when particular use cases, user needs and specific data of a domain are considered in detail. However, we believe that addressing these challenges is not only the key to enable Linked Data consumption in that particular domain, but also in other scientific domains by adopting the solutions.

References

1. Bemers-Lee T (2006) Linked data—design issues. http://www.w3.org/DesignIssues/LinkedData.html. Accessed 30 Apr 2015
2. Boland K, Ritze D, Eckert K, Mathiak B (2012) Identifying references to datasets in publications. In: Zaphiris P, Buchanan G, Rasmussen E, Loizides F (eds) Proceedings of the second international conference on theory and practice of digital libraries (TDPL 2012), LNCS 7489. Springer, Berlin, Heidelberg, pp 150-161
3. Bosch T, Zapilko B, Wackerow J, Gregory A (2013) Towards the discovery of person-level data: reuse of vocabularies and related use cases. In: Proceedings of SemStats 2013 collocated with ISWC 2013
4. Gottron T, Hachenberg C, Harth A, Zapilko B (2011) Towards a Semantic Data Library for the Social Sciences. In: Proceedings of SDA 2011, CEUR, vol 801
5. Heath T, Bizer, C (2011) Linked data: evolving the Web into a Global Data Space. Morgan & Claypool, San Rafael
6. Kommission Zukunft der Informationsinfrastruktur (2011) Gesamtkonzept für die Informationsinfrastruktur in Deutschland, http://www.leibniz-gemeinschaft.de/fileadmin/user_upload/downloads/Infrastruktur/KII_Gesamtkonzept.pdf
7. Ngonga Ngomo A-C, Auer S (2011) LIMES—a time-efficient approach for large-scale link discovery on the web of data. In: Proceedings of IJCAI
8. Schaible J, Gottron T, Scherp A (2014) Survey on common strategies of vocabulary reuse in linked open data modeling. In: Presutti V, dAmato C, Gandon F, dAquin M, Staab S, Tordai A (eds) The semantic web: trends and challenges; 11th international conference, ESWC 2014, Springer, LNCS 8645, pp 457^-72
9. Sure Y, Bloehdom S, Haase P, Hartmann J, Oberle, D (2005) The SWRC ontology—Semantic web for research communities. In: Bento C, Cardoso A, Dias G (eds) Progress in artificial intelligence, LNCS 3808. Springer, Berlin, pp 218-231. doi:10.1007/ 11595014_22
10. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Discovering and maintaining links on the web of data. In: Bernstein A, Karger D, Heath T, Feigenbaum L, Maynard D, Motta E, Thirunarayan K (eds) Proceedings of the 8th international Semantic Web Conference (ISWC '09), Springer, pp 650-665. doi: 10.1007/978-3- 642-04930-9_41
11. Wissenschaftsrat (2011) Empfehlungen zu Forschungsinfrastrukturen. Geschäftsstelle des WR
12. Zapilko B, Schaible J, Mayr P, Mathiak B (2013) TheSoz: a SKOS representation of the thesaurus for the social sciences. Semantic Web 4(3):257-263. doi:10.3233/SW-2012-0081
13. Zapilko B (2014) Methods for matching of linked open social science data. Dissertation, Universität Mannheim